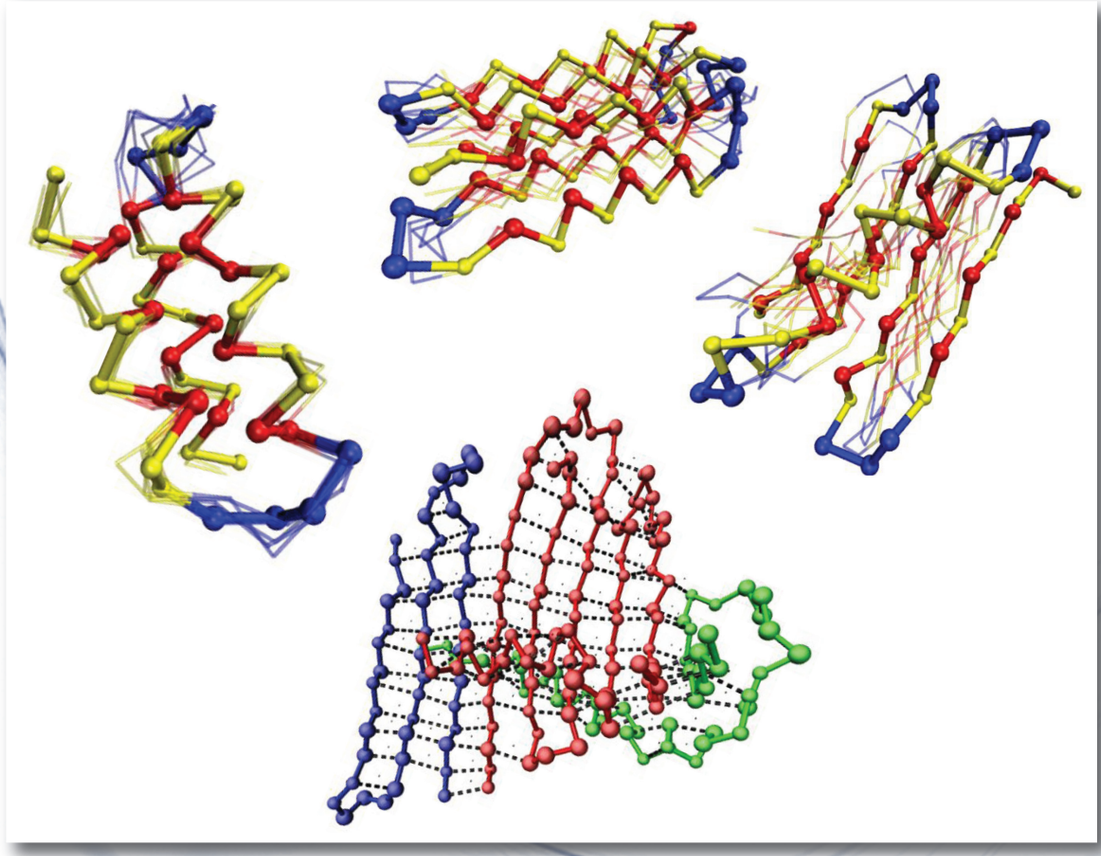
21 September 2013 Volume 139 Number 11

AIP | The Journal of Chemical Physics



80th Anniversary

jcp.aip.org

Sketching protein aggregation with a physics-based toy model

Marta Enciso and Antonio Rey

Citation: J. Chem. Phys. **139**, 115101 (2013); doi: 10.1063/1.4820793
View online: http://dx.doi.org/10.1063/1.4820793
View Table of Contents: http://jcp.aip.org/resource/1/JCPSA6/v139/i11
Published by the AIP Publishing LLC.

Sketching protein aggregation with a physics-based toy model

Marta Enciso[1] and Antonio Rey[2]
[1]*Institut für Mathematik, Freie Universität, D-14195 Berlin, Germany*
[2]*Departamento de Química Física I, Facultad de Ciencias Químicas, Universidad Complutense, E-28040 Madrid, Spain*

(Received 20 June 2013; accepted 21 August 2013; published online 16 September 2013)

We explore the applicability of a single-bead coarse-grained molecular model to describe the competition between protein folding and aggregation. We have designed very simple and regular sequences, based on our previous studies on peptide aggregation, that successfully fold into the three main protein structural families (all-α, all-β, and $\alpha + \beta$). Thanks to equilibrium computer simulations, we evaluate how temperature and concentration promote aggregation. Aggregates have been obtained for all the amino acid sequences considered, showing that this process is common to all proteins, as previously stated. However, each structural family presents particular characteristics that can be related to its specific balance between hydrogen bond and hydrophobic interactions. The model is very simple and has limitations, yet it is able to reproduce both the cooperative folding of isolated polypeptide chains with regular sequences and the formation of different types of aggregates at high concentrations. © *2013 AIP Publishing LLC*. [http://dx.doi.org/10.1063/1.4820793]

I. INTRODUCTION

Protein aggregation has cornered scientific research in the last decades, becoming a hot topic in the field. Thanks to the efforts of C. Dobson and other researchers, it has been found that aggregation, eventually leading to the formation of amyloid fibrils, is not limited to a relatively small number of proteins, but represents a generic feature of the polypeptide chains.[1,2] In fact, these fibril aggregates from different proteins share many characteristics, such as a common "cross-β" structure[3] formed by the stacking of β-sheets,[4] which can self-assemble without the need of other components. The core structure is stabilized by interactions involving the polypeptide main chain, particularly backbone hydrogen bonds. The competition between folding and aggregation can be understood in terms of the balance between specific, sequence dependent hydrophobic interactions (leading to a particular native state) and generic hydrogen bonds that result in a common β-type structure.[5]

The toughness of experimental aggregation research is broadly known. Studies of both disease-associated fibrils and of those formed from other proteins have enabled many of the features of these structures to be defined.[6,7] Although the detailed structure of amyloid fibrils resulting from peptide aggregation has been solved,[8,9] no complete protein aggregate in this state has yet been determined in atomic detail. Therefore, it is a field where simulations can be particularly useful. The main difficulty therein is the large size and timescales of the full aggregation process, well beyond standard full atom molecular dynamics simulations; these can be nevertheless applied to specific aspects, such as early stages of aggregation[10,11] or the simulation of fibril structures already formed.[12]

Coarse-grained simulations seem, then, a key option.[13] This strategy reduces the level of resolution of the system description, using one or a few interaction beads per amino acid and implicit solvent. Depending on the number of beads per amino acid, the complexity of the coarse-grained description varies and, with it, both its scope and limitations. Descriptions with many beads per amino acid have shown a remarkable success in the simulation of fibril nucleation and growth,[14,15] or in very recent cases also in the early stages of aggregation.[16] Models which use in their representations the heavy atoms in the protein backbone and one single bead for the side chain have been also successful in the study of peptide aggregation,[17] even trying to consider the competence between ordered ("folded") and disordered ("aggregated") situations.[18] Simpler, one-bead models (either lattice or off-lattice) have also been used, providing some insight on some particular aspects of protein aggregation, but mainly focused in peptide simulations[19,20] due to the inherent drawbacks of such simple models.[13]

The work we present here is framed among these latter models, as we aim to describe aggregation using a one-bead coarse-grained model. The main difference with previous approaches is the fact that we simulate full proteins, i.e., complete polypeptide sequences that can also fold into recognizable tertiary structures, without any additional information towards this state, at very dilute concentrations (single chain simulations). The simulation of full proteins using this level of resolution is a difficult goal on its own. Strategies using this simple level of resolution usually need to rely on *a priori* structure-based information. The best known example is the so-called Gō models, in which the interaction potential is defined in terms of the interactions that are present in the native state.[21] For this very reason, these models are not suitable for the study of non-native configurations such as aggregates, although some strategies like symmetrized or "colored" Gō potentials have been suggested.[22–25]

The possibility of using only realistic driving forces (that is, a complete removal of a reference folded structure) is tackled by the purely physics-based coarse-grained

potentials, which nevertheless have just proved their applicability in the case of sequenceless peptides.[26,27] An intermediate approach lies in the Sorenson-like potentials;[28,29] they use physics-based interactions for the calculation of hydrophobic interactions, but the local geometry (either helical, turn-like, or extended) is set *a priori* according to the desired secondary structure elements. Therefore, there is still a structural bias towards a reference native configuration.

Thanks to a single-bead coarse-grained potential we have recently developed,[30] we have even removed this latter constraint. Our model is based on an off-lattice α-carbon representation of the polypeptide chain and is only described by physics-based driving forces, namely hydrogen bonds and hydrophobic interactions. In a recent previous work we succeeded in simulating helical and β-type peptides with simplified sequences and no structural *a priori* information, presenting an aggregation study for these systems as a function of temperature and concentration.[30] In that work, the modeling effort was focused on an efficient description of secondary structure elements. Here, our aim is not only to reproduce the correct folding of a realistic tertiary structure in the very dilute regime, but also the aggregation of full proteins at high concentrations. The possibility to do both with a single bead per residue model, and without employing any reference towards a given ordered structure, other than the chain sequence itself, is what makes this study new and particularly appealing.

Our protein design pursues the stabilization of the three main protein folds, i.e., all-α, all-β, and $\alpha + \beta$. It has been suggested that general physicochemical principles such as the hydrophobic/polar patterning and the formation of hydrogen bonds are the key agents in the formation of secondary and tertiary structure.[5,31] The design of our sequences has followed these rules by using just a three-letter alphabet (hydrophobic, polar, and neutral amino acids) and very simple and regular sequences. The resulting proteins have been used afterwards to carry out a full set of temperature and concentration-dependent simulations. With them we have analyzed all these factors, building a schematic structural phase diagram where we show the folding/aggregation properties of our systems, depending on the chosen sequence, in terms of temperature and concentration.

II. METHODS

We describe here the main technical details that we have used to compute our data. First, we define the coarse-grained resolution that we have employed, as well as the main features of the physics-based potential (further details can be found in Refs. 30 and 32). Then, we describe the simulation conditions considered in this work.

A. System description and interaction potential

In our model, amino acids are described by one interacting bead, placed at the α-carbon position; beads that belong to the same polypeptide chain are connected through virtual bond vectors with a fixed length of 3.8 Å, corresponding to a *trans* peptide bond. The flexibility of the chain has been modeled according to the general characteristics of real proteins; then, the angle between three consecutive beads cannot be smaller than 65° nor greater than 150°; beads cannot be interpenetrated, so a hard-sphere bead-bead potential applies at distances smaller than 4.0 Å (a value that takes into account that our model beads represent the residues in the chain, and not just a single α-carbon), as previously optimized.[33,34]

The interaction potential that we have used here is based on this off-lattice α-carbon representation and includes the two main driving forces in protein systems: hydrogen bonds and hydrophobic interactions. Detailed descriptions can be found in Refs. 30 and 32.

The global energy is computed as follows:

$$E = \omega^{hb} E^{hb} + \omega^{hp} E^{hp} + \omega^{stiff} E^{stiff}. \qquad (1)$$

In this expression we can find three terms, each of them including a certain weighting factor, ω. The first term models the formation of a hydrogen bond between a certain pair of residues i and j (with $i < j$), where $j > i + 2$ and $j \neq i + 4$. Given one of these pairs, we define some auxiliary vectors, sketched in Figure 1: vector $\mathbf{r}_{i,j}$ connects the residues; vectors $\mathbf{h}_i$ and $\mathbf{h}_j$ are unitary vectors perpendicular to the plane defined by a residue and its preceding and following neighbors. Then, the hydrogen bond energy is calculated in two steps. The first one involves the computation of some geometrical quantities that depend on those vectors (the length of vector $\mathbf{r}_{i,j}$; the relative orientation between $\mathbf{h}_i$ and $\mathbf{h}_j$; and the relative orientation between those $\mathbf{h}$ vectors and $\mathbf{r}_{i,j}$). According to the geometry of real secondary structure elements, a hydrogen bond is formed only if the values of these quantities lie within certain limits; in that case we apply a stepwise potential. Acceptable ranges and potential strength differ depending on the kind of hydrogen bond (either local or

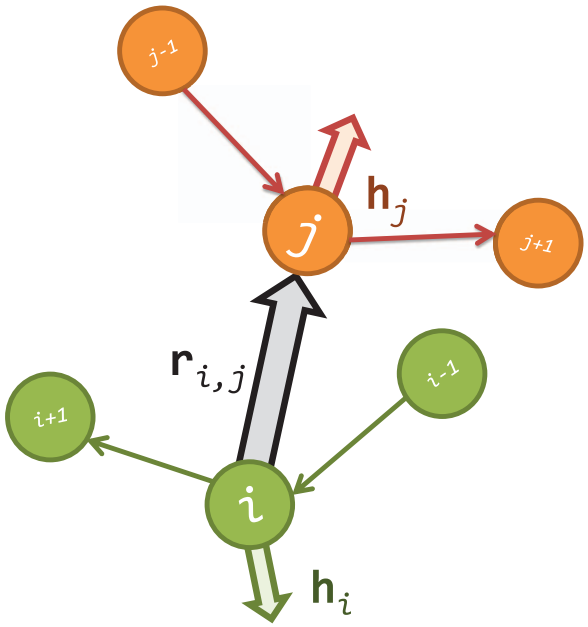


FIG. 1. Schematic representation of the vectors needed to compute the hydrogen bond interaction, according to our potential definition. Two protein segments have been drawn, colored in green and orange. Vectors $\mathbf{h}$ are represented by colored arrows and the connecting vector $\mathbf{r}_{i,j}$ is colored in black.

non-local, as a function of the separation between the interacting beads along the sequence); the exact definition of the geometrical quantities as well as the acceptable ranges can be found in Ref. 32.

The hydrophobic interaction uses a 12–10 Lennard-Jones definition that is attractive in the case of an interaction between two hydrophobic residues and repulsive otherwise,

$$E_{i,j}^{hp} = S_1 \left[\left(\frac{\sigma_{int}}{|\mathbf{r}_{i,j}|} \right)^{12} - S_2 \left(\frac{\sigma_{int}}{|\mathbf{r}_{i,j}|} \right)^{10} \right], \qquad (2)$$

where $|r_{i,j}|$ is the distance between residues i and j. S_1, S_2, and σ_{int} are optimized parameters; a detailed description of this interaction and its parameters can be found in Ref. 30. We only distinguish among three types of residues: hydrophobic (H), polar (P), and neutral (N), following earlier works of Head-Gordon and co-workers.[28] It represents a very simple form of a mean field interaction potential, which results in an effective attraction between H residues and an effective repulsion between P residues, implicitly mimicking the effect of the aqueous solvent.

We have also included a stiffness term that, coupled in occasions to the hydrophobic interaction previously described, avoids a too strong hydrophobic collapse by controlling the chain flexibility. There are multiple ways to define this term. Most of them insert a geometrical bias to favor a certain secondary structure element, modifying the applied functional form accordingly;[28,29] in our definition we apply the same functional form in all cases, introducing no preferential geometry. This definition controls the torsional angle of each residue i, ϕ_i, which is defined by four consecutive beads in the chain,

$$E_i^{stiff} = \omega_t \left[0.5 \left(1 + \cos 3\phi_i \right) - 1 \right]. \qquad (3)$$

Then, it equally favors helical conformations (in this model, with $\phi_i = \pm 60^\circ$) and extended ones ($\phi_i = \pm 180^\circ$). We have included the weighting factor ω_t to model the additional flexibility of loops; then, $\omega_t = 1.0$ in the case of structured regions (whatever secondary structure is formed) and drops to $\omega_t = 0.4$ in the case of loops, reducing the rigidity of this part of the chain but without any impact on the preferred conformation.

The aforementioned energetic terms in Eq. (1) are added up to build the final energetic interaction according to their weighting factors, ω. These values have been optimized to stabilize secondary structure elements and provide reasonable folding properties in the case of peptides.[30] These values are the following: $\omega^{hb} = 9.5$, $\omega^{stiff} = 7.0$, and $\omega^{hp} = 6.5$. Note that our simulations are performed in reduced units, defined in terms of a certain reference temperature, T_{ref}: $T^* = T_{real}/T_{ref}$ and $E^* = E_{real}/(k_B T_{ref})$.

B. Simulation details

We have used a Replica Exchange Monte Carlo (REMC) in-house simulation program, parallelized with OpenMP for higher performance over multi-core processors. We have performed single-chain and multi-chain numerical experiments (in this latter case, using periodic boundary conditions in a simulation box), where each of our simulations spans 24–40 temperatures (from 1.8 to 2.9 in reduced units). The largest temperatures correspond to a situation where individual, unfolded conformations appear for every chain in the simulation, and therefore contribute to avoid that the simulations get trapped in local minima. The lowest temperatures correspond to the folded conformations alone, in the case of individual chains, or to well defined energetic and structural situations for the multi chain systems, representing our aggregates (see below). All the simulations start from a completely extended conformation for each chain and consist of 8×10^6 Monte Carlo cycles at every temperature after 3×10^6 equilibration cycles. In each cycle, every bead of the system is subjected to a trial Monte Carlo move. We have also included rigid shifts and rotations of individual chains, or a group of them, to allow for the diffusion of some chains relative to the others within the simulation box. The standard characteristics of a replica exchange simulation methodology (number of temperatures and their values, frequency of replica exchange trials and their acceptance ratio, the proper travel of the replicas along the different temperatures, etc.) have been tested in our group for different interaction models of single and multiple chain systems in recent years,[30,33–35] and they are also checked here to warrant a proper sampling, which allows for an accurate description of the equilibrium properties for the studied systems. Moreover, for each system, three or five independent REMC runs have been carried out, in order to provide statistically meaningful results. The statistical coincidence of the results from the individual runs is an additional test of the quality of the sampling procedure. The results reported here correspond to the average on the independent simulations for every system.

In the case of multichain numerical experiments, concentration has been modeled by using four chains in simulation boxes of different sizes, depending on the length of the chains and the desired concentration. The box size is large enough to warrant that a chain never interacts with itself through the minimum image convention. Although simulations with larger number of chains have been reported in the literature in the study of peptide aggregation and fibrillization,[17,36] four chains are a reasonable number to study the competition between full protein folding and aggregation, resulting in a system feasible to be fully analyzed with modest computational resources. We have also performed additional simulations with just two chains per box, leading to similar results, at the level described here. The data we present here belong to the four chain simulations. We have modeled systems in the range from 0.1 to 5.0 residue moles/L. We use these unfrequent concentration units (moles of amino acids per unit of volume) since our polypeptide chains have different number of residues depending on their sequences. This way, we have a similar concentration variation for the three simulated proteins. Taking into account the chain lengths (see below), our concentrations would be in the approximate range from 2 to 100 mM (in protein moles). Note, however, that the numerical values of the simulated concentrations exemplify the variation analyzed in this work, but they do not try to quantitatively reflect a real experimental concentration.

III. RESULTS AND DISCUSSION

In this section, we introduce the simplified proteins that, in combination with the single-bead coarse-grained model described above, help us to grasp the main folding and aggregation characteristics of proteins under different temperature and concentration environments. In order to be as general as possible, our targets belong to the three main protein folds: α, β, and $\alpha + \beta$. For the sake of clarity, this section is divided in three parts, each of them matching a different fold. We want to stress that our study does not refer to specific real proteins, but tries to constitute an analysis of the physics underlying the protein universe. This is the reason why we use three different sequences, leading to different stable structural families. The comparison among their results aims to make this study as general as possible, always as much as our simple model permits.

For each structural family, we have designed a protein sequence that stabilizes the desired global shape. Given the simplicity of our model, and particularly the usage of just a three-letter alphabet for the sequence, our aim has not been the simulation of a real protein, as stated above, but the analysis of a simplified version built with simple and very regular sequences. Then, we start by carrying out single chain simulations to examine the folding properties of the designed protein. Thanks to them we check if general protein features are properly modeled, such as a well defined native state and a cooperative, two-state folding transition. Afterwards, we perform extensive multichain simulations at different temperatures and concentrations, and analyze the aggregation propensities and the characteristics of the different transitions along the configurational space. As a result, we have computed for every protein a sketched structural phase diagram in which we summarize all this information.

A. All-α protein

All-α proteins are characterized by the presence of several α-helices packed against one another. This packing is driven by the presence of a certain hydrophobic core which needs to be preserved from the solvent. Following this rule of thumb, we have designed a compatible sequence composed by helical fragments connected by short turns.

The design of a helical protein has been relatively straightforward, as we already had some information on how helical fragments associate at low or moderate concentrations with our coarse-grained model.[30] In our previous work with peptides, we found that α-helices mainly stabilize three helix bundles, which is also a very broad feature in real proteins, although four chain bundles have been also considered in other works.[18] Each helical segment (13 residues long, matching the average length of this secondary structure element) consists in a periodic repetition of polar (P) and hydrophobic (H) residues in the following pattern: $(\mathrm{HPPHHPP})_n$;[29,30] these fragments are connected in our protein sequence by turns (formed by four residues) that have neutral (N) character. Therefore, our resulting sequence is PPHPPHHPPHPPHNNNNPHPPHHPPHPPHHNNNNPPHHPPHPPHHPP. The full modeled chain has a length of 47 residues, which comes from the design criteria mentioned above.

We have first analyzed the thermal stability of our simplified protein by single chain REMC simulations, as a function of the temperature alone. From the resulting energy fluctuations, we have computed the heat capacity at the different temperatures sampled, as a peak in this thermodynamic property indicates an energetic transition, which in turn implies also a structural one. In this particular case, we have obtained the heat capacity curve of Figure 2(a). It shows a single and relatively narrow peak, which indicates that the all-α protein has a well-defined and quite cooperative folding/unfolding transition. We have explicitly checked that this is indeed the case by computing the energy histogram from

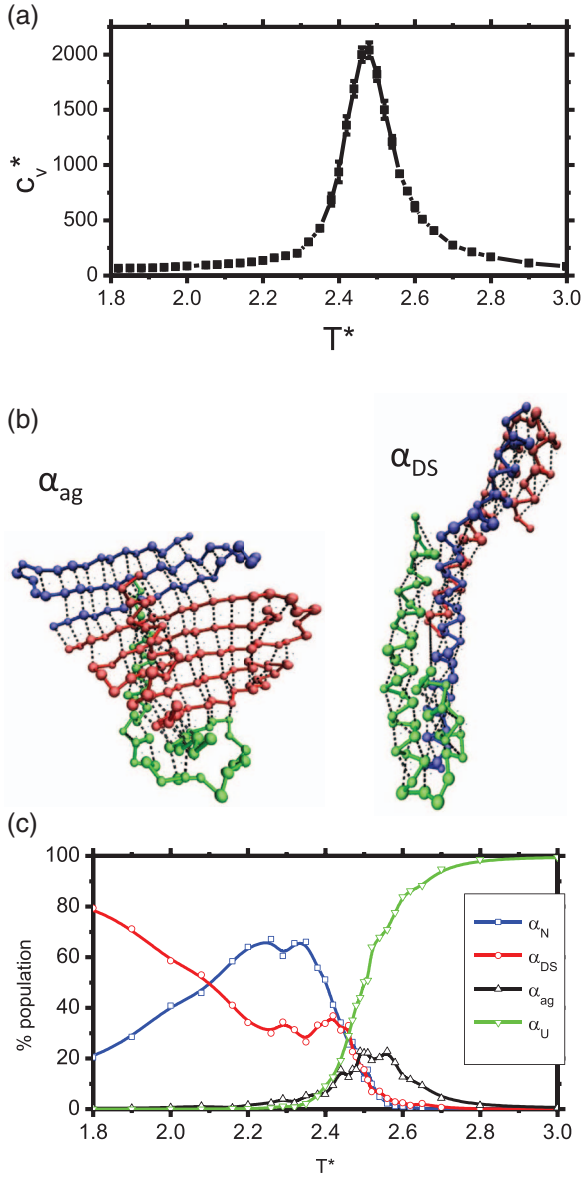


FIG. 2. (a) Heat capacity curve versus temperature for a single chain with helical sequence. (b) Representative configurations of the observed multichain clusters in our helical system: α_{ag} (aggregate) and α_{DS} (domain swapping). In all multichain structures each chain presents a different color. The dotted lines correspond to the hydrogen bonds in the model. (c) Population of the different structural clusters of our helical system at 5.0 residue moles/L.

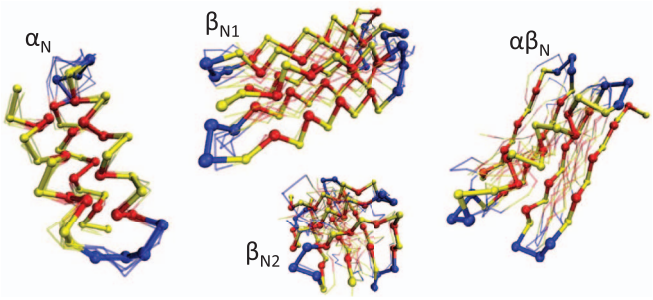


FIG. 3. Clusters at the native states of all-α (α_N), all-β (β_N), and $\alpha + \beta$ ($\alpha\beta_N$) proteins. In these schemes, hydrophobic residues (H) have been colored in red, polars (P) in yellow, and neutrals (N) in blue.

the simulations at different temperatures. The results (shown in Figure S1 in the supplementary material[48]) show bimodal distributions that switch from a peak at low energies corresponding to the folded state (with its thermal fluctuations) below the transition temperature, i.e., the temperature corresponding to the peak of the heat capacity curve, to a different peak at higher energies which becomes progressively more populated as the temperature is increased. The presence of a well defined minimum between these two maxima when both are present at intermediate temperatures is the signature of a cooperative transition. In addition, these histograms also show the lack of intermediates with relevant populations along the folding/unfolding transition.

In Figure 3 we show the native state of the all-α sequence, labeled α_N. We observe that the three α-helices are correctly formed; each of them is packed against the rest, burying the hydrophobic residues (in red). The small fluctuations around this structure, shown by the thin line structures in Figure 3, are also reflected in the small value of the heat capacity at low temperatures, indicating that this structure is stable and well defined, as it would correspond to a native state. A deeper structural analysis of the simulation results confirms that the three helix bundle loses its folded structure at the transition temperature without any populated thermodynamic intermediates (see Figure S1 in the supplementary material[48]). Therefore, we have succeeded in the design of this protein, finding one single folded conformation that exhibits a cooperative two-state folding process.

Then, we have explored the aggregation properties of our helical protein in terms of concentration and temperature. As mentioned in the Introduction, it is well known that many proteins (regardless of their folded shape) present frequent interchain interactions and even aggregation at high concentration conditions.[37] We have simulated five systems with concentrations of 0.2, 0.5, 1.0, 2.0, and 5.0 residue moles/L, and a full temperature range in the REMC scheme for every case. From the results we have got for each of these systems at every temperature, we have performed a clustering analysis based on the different structures we have observed along the simulations (which are identified through the values of different properties, such as the type and strength of their interactions, the radius of gyration, etc.). A further explanation of this procedure is included in the supplementary material.[48] Among the ensemble of conformations, we have recognized three main structural motives, apart from unfolded configurations (labeled α_U in what follows). The first one is the helical native state (α_N, see Figure 3) in which every chain inside the simulation box is independently folded, presenting the characteristic three-bundle shape that we also find in single-chain simulations below the transition temperature. The second cluster consists in domain-swapped helical bundles (α_{DS}). It is a specific type of aggregation, in which two or more chains interchange the packing of equivalent domains (in this case, individual helices) from one protein to another. The existence of domain swapping in different types of protein structures and its influence in protein aggregation and disease have been previously reviewed.[38] As an example of our results, we show in Figure 2(b) one of the possible domain-swapping configurations where three chains are involved. Note that a high helicity content is kept (i.e., the local geometry still resembles the native one). The hydrogen bonds found by the model in this configuration, represented by dotted lines in the figure, are still local, stabilizing the helical conformation of the different chains, which interact among them mainly through attractive hydrophobic interactions. The third motif in our clustering analysis is the β-type aggregate (α_{ag}): the most characteristic feature of these configurations is the presence of many long range intrachain and interchain hydrogen bonds, shown with dotted lines, forming sheet-like structures such as the one in Figure 2(b). In this latter case, the local geometry (mainly extended) is completely different from the native one (helical) and resembles the β-type configurations that have been found in real aggregated proteins of any sequence.[39]

After identifying these clusters, we have analyzed the conditions (i.e., temperature and concentration) at which they are found. As an example, we show in Figure 2(c) this analysis for the concentration 5.0 residue moles/L. If we look at the population balance at low temperatures, we find that swapped-domains (α_{DS}) are more abundant, although a smaller population (around 20%) of native configurations (α_N) is also present. Do not forget that these results correspond to the highest concentration we have simulated. At intermediate temperatures there is an inversion in this balance, prevailing α_N. At $T^* \simeq 2.4$ there is a change in the trend: the population of unfolded configurations (α_U) begins to raise, as well as a certain population of aggregated structures (α_{ag}), which remains significant up to $T^* \simeq 2.6$.

We have repeated this analysis for all the simulated concentrations. We have identified different stability regions for each type of the configurations described above, and plotted this information into the sketched structural phase diagram of Figure 4, where we have considered that a certain structure is present if its relative population at the evaluated conditions is equal to or greater than 20%. Note that the separation among different regions is marked with straight lines instead of a realistic boundary; this is a result of the lack of knowledge on these boundaries, especially along the vertical axis, as only some individual concentrations, indicated above, have been sampled. The temperature axis is better determined, since the number of temperatures used in the REMC procedure, already described, is quite large (with individual values indicated by the positions of the symbols in Figure 2(c)).

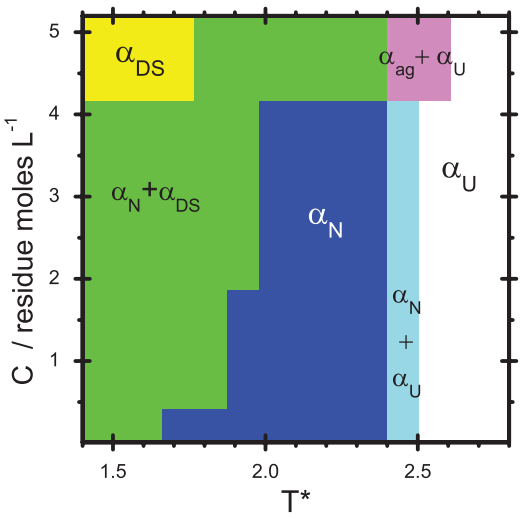


FIG. 4. Sketched structural phase diagram corresponding to our helical protein under different temperature and concentration conditions. Regions where one or more structures are predominant (i.e., population greater than 20%) have been marked with colors.

If we start from the low temperature region in Figure 4, domain-swapped conformations are always present to some extent, but concentration determines the details of the structural scenario. High concentrations favor only α_{DS}; in more diluted systems, α_{DS} coexists with independent helix bundles (α_N); the coexistence region is shifted towards lower temperature values when the concentration becomes lower, showing a reduction in the stability of this DS multichain structure due to the less frequent interchain interactions when the system becomes more diluted. Also, although the way we have sketched the phase diagram only uses the criterion of a minimum population (20%), the region where α_{DS} and α_N coexist is not uniform, in the sense that the relative populations between these two situations vary along the common area in the diagram. At the lowest concentration represented in Figure 4, the population of α_{DS} is barely 20%, but the population of α_N adds up to the remaining 80% below the unfolding temperature. The fact that α_{DS} appears with a significant population at low temperatures, even at this low concentrations, may be related to the simplicity of the model. A very low temperature is equivalent to an energy minimization of the simulated system. Therefore, situations as α_{DS} with a large number of interchain contacts, which imply the formation of attractive hydrophobic interactions, are favored in these conditions, even if the energetic difference with a configuration of individual folded chains is very small. When the temperature increases, this small additional stability is not enough to deal with the thermal fluctuations, and α_N is essentially the only populated state in this dilute regime. Probably, one should not extend the use of this type of very simple models to temperatures too low in comparison to the unfolding temperature of the isolated chains, to avoid this type of undesired effect.[32]

In the less concentrated systems we have considered for this protein (from 0.2 to 2.0 residue moles/L), the stable configuration at intermediate temperatures is α_N, that is, the helical native state. As it is an isolated, single-chain protein structure, its stability does not depend on concentration; therefore, its unfolded transition temperature is always found at $T^* = 2.5$ (similarly to the transition temperature in single-chain simulations, see Figure 2(a)). In the highest concentrated case (5.0 residue moles/L), however, multichain structures are found at every temperature below the transition towards the unfolded state, α_U; α_{DS} coexists with α_N until $T^* = 2.4$; slightly above that temperature we find a region where β-type structures, α_{ag}, coexist with unfolded configurations. Thus, according to our model, β-aggregates are found around the transition temperature even in the case of helical sequences, if concentration is sufficiently large.

B. All-β protein

In all-β proteins the secondary structure is composed by β-sheets (i.e., extended segments which interact through nonlocal backbone hydrogen bonds). These sheets are packed so that their hydrophobic faces become buried in the inner part of the overall structure. Again, we have used simple hydrophobic/polar patterns to build our all-β toy protein, relying on our previous results on independent peptides.[30] According to them, six β-type fragments were stabilized through a $\beta_3 + \beta_3$ sandwich. Each of these β fragments (9 residues long) is modeled in the three letter alphabet by an alternating succession of hydrophobic and polar segments, $(HP)_n$.[29,30] In this work we have linked them by turns of three neutral residues, resulting in the following sequence, with 57 residues: PHPHPHPHPNNNPHPHPHPHPNNNPHPHPHPHPNNNPHPHPHPHPNNNHPHPHPHPP. Again, it corresponds to the possibility of reproducing, in the frame of a single covalent sequence, the packing of H and P residues we found with the same interaction potential using small peptides.[30]

As before, we have analyzed the stability of the resulting protein by single-chain REMC simulations. Similarly to the all-α protein, we have found a heat capacity curve (see Figure 5(a)) with a single and very narrow peak at $T^*_m = 2.76$, confirming the existence of a single energetic transition in the system. The histograms for the energy as a function of temperature, shown in Figure S1 in the supplementary material,[48] confirm that, from an energetic point of view, the model shows a two-state, highly cooperative folding/unfolding transition, without populated intermediates. The clustering structural analysis, however, reveals the presence of two different compact structures at low temperatures, which we have named β_{N1} and β_{N2}, whose populations appear in the proportion 85:15; these clusters are represented in the central part of Figure 3. Both share a $\beta_3 + \beta_3$ stacking, but differ in the relative arrangement of the upper and lower layers: parallel in β_{N1} and perpendicular in β_{N2}. This finding is not surprising, as we had found also two similar β-type structures in our peptide simulations.[30] As a matter of fact, the packing of the two sheets relies in the hydrophobic interactions among the different residues. The subtle differences existing in real proteins are, in our case, dimmed by the use of an extremely regular sequence and model, where all the β segments are identical and defined with just two different types of residues.

The lack of side chains in our model, which also broadens the packing possibilities, can be also blamed for this situation. While it represents a clear caveat of our model, at least for the simulation of all-β structures, the results are still interesting. As we have seen, the heat capacity curve exhibits a single and relatively sharp peak at the transition temperature, with an essentially flat tail without any relevant feature at low temperatures, which means that these two structures are energetically indistinguishable from each other in our model. In addition, the population of β_{N1} at low temperatures, which exceeds to that of β_{N2} by a factor larger than 5, makes us think that our model and design strategy is essentially correct also for the all-β protein. Although the toy model, by definition, has evidently room for a lot of improvement, the very large population of β_{N1} in the folded state of this sequence, which stays much larger than the population of β_{N2} up to the transition temperature, permits it to be considered as a reasonable, although not perfect, candidate for the type of study undertaken in this work.

The evaluation of the aggregation properties of this all-β protein has been performed in similar terms as the all-α case, carrying out multichain simulations at concentrations of 0.1, 0.2, 0.5, 1.0, and 2.0 residue moles/L. This means a slight shift of the concentration scale towards lower values, in comparison to the concentration range we have used in the all-α protein. We have started with a structural clustering, as described above and in the supplementary material,[48] and followed with the evaluation of each clusters' population at every concentration and temperature. Regarding the clustering itself, we have observed two main types of compact structures: native ones (mostly β_{N1} with a minor population of β_{N2}) and aggregated ones (that we present in Figure 5(b)). Note that β_{ag} is very similar to α_{ag} in Figure 2(b), with a full network of intra- and inter-chain hydrogen bonds, supporting the idea of common structural features for the aggregates resulting from different proteins.[2]

The cluster populations vary at different environment conditions, as shown in the structural phase diagram of Figure 5(c). We can observe a very different situation compared to the all-α case, since aggregated structures are much more frequent in the all-β protein. This is the reason why it is not necessary to increase the concentration up to 5.0 residue moles/L for this protein, as it was in the all-α protein. If we observe intermediate and high concentration values in Figure 5(c), aggregates are the leading configurations at low and intermediate temperatures. At higher temperatures, every regular structure is lost, mainly finding unstructured configurations, β_U. At lower concentration levels (0.1 residue moles/L), low temperatures still lead to β_{ag} configurations, although this fact can be related to the "energy minimization conditions" imposed by these low temperatures, as explained in the all-α results. However, intermediate temperatures show a certain population of native isolated structures, either β_{N1} or β_{N2}, always with a majority of the former. Above that temperature segment, the system evolves to an unstructured situation. The competition between folding and aggregation for this all-β protein is, therefore, clearly shifted by sequence towards aggregation. When compared to the all-α protein, the trend of the all-β protein towards aggregation in our model is so large

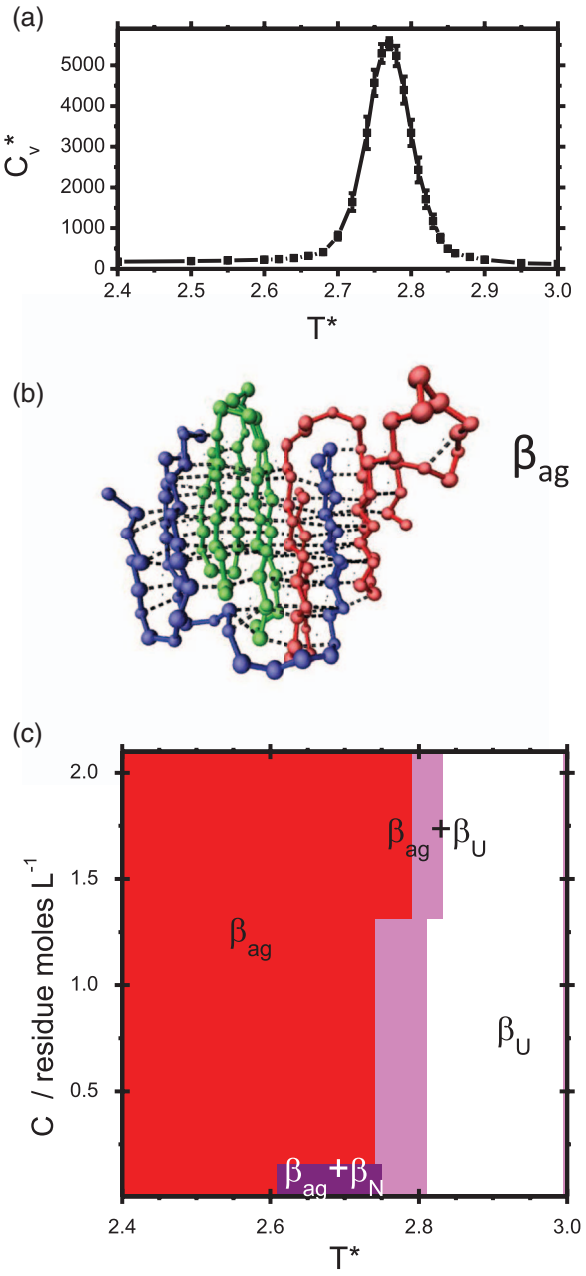


FIG. 5. Results for the all-β protein. (a) Heat capacity curve versus temperature for a single chain. (b) Schematic picture of the multichain structure β_{ag}. The dotted lines correspond to the hydrogen bonds in the model. (c) Sketched structural phase diagram.

that the single chain, folded conformation is only found at very low concentrations.

C. $\alpha + \beta$ protein

Proteins with $\alpha + \beta$ structures combine helical and extended segments in such a way that most of the hydrophobic residues avoid solvent exposure. Following our sequence design strategy, we have initially simulated peptides with helical and β-prone sequences in different proportions, finding that the most stable multisegment structure is formed (in our model) with four β segments which associate into a sheet, and one helix packed against one of their surfaces. Thus, we have linked these

segments by loops of three neutral residues, resulting the sequence PHPHPHPHPNNNPHPHPHPHPNNNPPHHPPH-PPHHPPNNNPHPHPHPHPNNNPHPHPHPHP, that is, two β-strands forming a β-hairpin, the α-helix, and two more β-strands in a new hairpin, in a total of 61 residues. It is the smallest $\alpha + \beta$ sequence we have found with our model which is adequate for a correct folding at very low concentrations.

The stability of this protein has been evaluated through single chain simulations, like in the aforementioned proteins. The heat capacity curve as a function of temperature presents one narrow peak at $T_m^* = 2.56$ (see Figure 6(a)): it matches a two-state transition between a low-temperature folded structure (labeled $\alpha\beta_N$ in Figure 3) and unstructured conformations at higher temperatures. We can observe in Figure 3 that the desired fold is obtained below the transition temperature: the four β-strands form a sheet, packed against an α-helix that buries its hydrophobic face. The characteristics of a two-state transition without intermediates have been also checked, as in the other two sequences, through the energy histograms as a function of temperature (see Figure S1 in the supplementary material[48]).

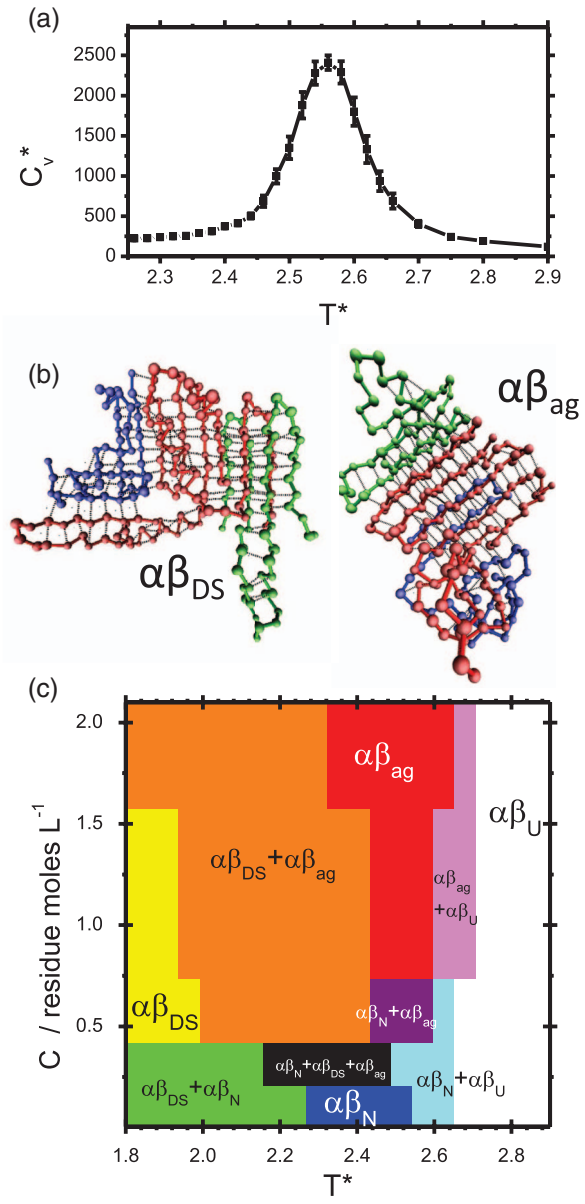


FIG. 6. Results of the $\alpha + \beta$ toy protein. (a) Heat capacity curve versus temperature for a single chain. (b) Schematic pictures of the structural multichain clusters in this system: $\alpha\beta_{DS}$ shows one of the possible structures with domain-swapping; $\alpha\beta_{ag}$ is the aggregated β-type multichain structure. The dotted lines correspond to the hydrogen bonds in the model. (c) Sketched structural phase diagram in different temperature and concentration conditions.

This toy protein has been also subject to multichain simulations in a concentration range from 0.1 to 2.0 residue moles/L, with the same individual values for concentration as in the all-β protein. We have identified three types of different structural clusters: domain swapped configurations (labeled $\alpha\beta_{DS}$ in Figure 6(b)), β-type aggregates ($\alpha\beta_{ag}$ in Figure 6(b)), and configurations of the system where every chain is in the native conformation ($\alpha\beta_N$). We have performed the usual clustering procedure in terms of temperature and concentration, building the structural phase diagram of Figure 6(c). We can observe there an increase in the complexity of the configurational space, as structures coexist in many different conditions. In general terms, we can say that domain-swapped structures ($\alpha\beta_{DS}$) are mainly found at low temperatures, either isolated or in combination with aggregates ($\alpha\beta_{ag}$) at high concentrations, or with native configurations ($\alpha\beta_N$) at lower concentrations. Aggregates are predominant at high concentrations and temperatures around the unfolding transition. In the same temperature range but at low concentrations, we can find isolated folded structures. Again, the trend towards aggregation for this protein, either in the form of domain swapped structures or as β-aggregates, is significantly larger than in the all-α protein, and similar to the all-β one.

IV. SUMMARY AND CONCLUSIONS

Single bead coarse-grained potentials are broadly used in protein folding because of its low computational cost, derived from their simplicity.[40] But this aspect also involves some drawbacks; the most important of them, according to our purposes, is the need of a reference structure to define the native state and/or local geometry, which hinders a realistic competition between folding and aggregation. We have tried to fill this gap by a careful design of a single-bead coarse-grained potential that does not rely on any folded structure reference, but on the use of a regular sequence. In a previous work,[30] we successfully applied it to peptide systems, where we used a three-letter hydrophobic alphabet to design peptides that, depending on their sequence, stabilized different secondary structures. Thanks to that knowledge, we have combined these secondary structure elements to build the complete proteins we have presented in Sec. III, and whose structural features in the native state are shown in Figure 3. We have designed three different proteins, each of them belonging to a different structural family. In all the cases, we have observed a neat two-state folding equilibrium, similar to the usual thermal response of real, single-domain globular proteins. The clustering analysis of the simulation results computed on single chain systems has revealed the key role of sequence in the formation of the native state: secondary structure elements are reproduced

as expected from the hydrophobic/polar patterning along the sequence, and they are packed in a way that enhances the burial of hydrophobic residues.

In the case of the all-α and $\alpha + \beta$ proteins, we have identified a unique native state, stable at low temperatures in single chain simulations. Regarding the all-β protein, we have detected two alternative folds that are equally stable, although the one pursued in our design is by far the most populated below the folding transition temperature. Given the outstanding regularity of the used sequence and the plainness of a three-letter alphabet, the presence of an alternative, though minor, folded conformation may be assigned to the lack of side chains in our model and of any particularities for each amino acid (both at the level of geometry and interactions) in the folding process. Anyhow, both structures comply with the initial expectations of our design, namely a $\beta_3 + \beta_3$ fold.

Therefore, we have shown in this work that a simple strategy based on the use of hydrophobic patterning and hydrogen bonds is enough to reproduce the main features of real proteins' folding. In fact, it has been suggested that these two actors may be even more important than residue-specific identities in the achievement of a certain secondary and tertiary structure.[5,31]

The competition between folding and aggregation has been examined through multichain simulations of the aforementioned toy proteins at different temperature and concentration conditions. We have identified different structural clusters and their stability regions, which have been plotted in the sketched structural diagrams of Figures 4, 5(c), and 6(c).

The helical protein can form two types of multichain structures, both of them also experimentally found in real helical proteins. We have found for this protein β-type aggregates stabilized by interchain hydrogen bonds, like α_{ag} in Figure 2(b), that resemble amyloids.[4] Besides, we have identified domain-swapped configurations (such as structure α_{DS} in Figure 2(b)), which are also thought to play some role in amyloid formation mechanisms.[41,42]

The phase diagram in Figure 4 shows the effect of environment conditions on this system. Then, β-aggregates are observed at very high concentrations (in the simulated scale) and only at intermediate temperatures, which can be related to the need of partially unfolded conformations so that the aggregation process may happen.[43] Domain-swapped configurations are present in a different stability range: although they are more favored at high concentrations, they can also be found to some extent even at mid to low concentrations, as long as temperature is low enough, which can be related to the additional interactions stabilizing these configurations, a number which is small but able to keep a significant population of domain-swapped configurations at low temperatures and concentrations. This has also been observed with other models and conditions, and at sensible temperatures (not too low) highlights the connection between domain-swapping and minimally frustrating driving forces.[44]

Our all-β protein is clearly more sensitive towards aggregation, as it could be inferred from its higher β-propensity.[45] The phase diagram of Figure 5(c) shows that β-aggregates (sharing similar structural features to α_{ag}) are present in our system even at low concentrations. As a matter of fact, the single chain folded structure is only stable for this protein in highly dilute conditions, according to our model. Thus, even though aggregation is shown to be a common process in all types of proteins, our results show that sequence still plays an important role in the different aggregation propensities, something which has already been realized.[46] Moreover, the folding characteristics of this sequence, when studied in isolated chain simulations, show that it is the most cooperative of the cases considered in this work, as shown by the narrowest heat capacity peak among those obtained here, and the very marked two-state character of the energy histograms along the thermal transition. Yet, it is the one which, according to our model, results more aggregation-prone. Although the presence of intermediates may favor aggregation,[47] something that had been also checked with simple simulation models,[23] our results show additional evidence that aggregation is also possible for two-state proteins. According to our model, this fact is related to the protein sequence and the compatibility between the hydrophobic interactions and the backbone hydrogen bonds, not only in the folded conformation but also in the domain-swapped configurations and in the β-type aggregates. In the latter, the simplicity of the model cannot produce results fully compatible with the experimentally observed cross-β structures. This is too much to ask for a model with a very simple sequence and interaction scheme, which is also devoid of side-chains. However, the structures shown in Figure 5(b) allow us to believe that the model could add up more chains in our β-aggregates, in case they were included into the simulated system.

In the case of the $\alpha + \beta$ protein, intermediate properties are expected between the other two proteins considered in this work, and this is what we get in our results. The structural phase diagram of Figure 6(c) shows a strong interplay among native configurations, swapped domains, and aggregated structures. Domain-swapping configurations are, like in the all-α protein, found at low temperatures all over the concentration range (alone or coexisting with either β-aggregates or the native state). Native conformations and aggregates share the mid-temperature slot; the competition between them is mediated by concentration, illustrating the well-known role of this factor in aggregation.[2] The combination of two different secondary structure elements in a single sequence introduces a higher sequence complexity and reduces its periodicity, resulting in a deeper interplay among structures in the configurational space and a more realistic view of how more complex proteins may behave.

In a nutshell, we have successfully found the basic hallmarks of protein folding and aggregation using a very simple approach in terms of sequences and the simulation model itself. It may appear that our main finding, the fact that the system shows a higher trend towards aggregation at high concentrations when the β content of the native state becomes larger, represents just an expected, not very interesting result. But we consider we have been able to relate this fact to very basic physical principles of the main interactions responsible for protein stability, mainly the adequate interplay between the backbone hydrogen bonds and the sequence dependent hydrophobic interactions. In addition, we have got these results and conclusions with a very simple simulation

model, which is able however to very reasonably fold a simple regular sequence towards a desired folded conformation without any additional information on this structure, something which is far from trivial at this level of simplification.

ACKNOWLEDGMENTS

This work was partially supported by the Spanish Ministerio de Ciencia e Innovación (Grant No. FIS2009-13364-C02-02), and by Comunidad Autónoma de Madrid (Grant No. S2009/PPQ-1551). M.E. acknowledges a former Scholarship from Spanish Ministerio de Educación (FPU program), and presently a Scholarship from the European Marie Curie program.

[1]F. Chiti, P. Webster, N. Taddei, A. Clark, M. Stefani, G. Ramponi, and C. M. Dobson, Proc. Natl. Acad. Sci. U.S.A. **96**, 3590 (1999).
[2]C. M. Dobson, Nature (London) **426**, 884 (2003).
[3]M. Sunde, L. C. Serpell, M. Bartlam, P. E. Fraser, M. B. Pepys, and C. C. F. Blake, J. Mol. Biol. **273**, 729 (1997).
[4]J. I. Guijarro, C. J. Morton, K. W. Plaxco, I. D. Campbell, and C. M. Dobson, J. Mol. Biol. **276**, 657 (1998).
[5]W. Anthony, T. P. Knowles, C. A. Waudby, M. Vendruscolo, C. Dobson, and M. Fitzpatrick, PLOS Comput. Biol. **7**, e1002169 (2011).
[6]R. Tycko, Annu. Rev. Phys. Chem. **62**, 279 (2011).
[7]N. Norlin, M. Hellberg, A. Filippov, A. A. Sousa, G. Gröbner, R. D. Leapman, N. Almqvist, and O. N. Antzutkin, J. Struct. Biol. **180**, 174 (2012).
[8]R. Nelson, M. R. Sawaya, M. Balbirnie, A. O. Madsen, C. Riekel, R. Grothe, and D. Eisenberg, Nature (London) **435**, 773 (2005).
[9]J.-P. Colletier, A. Laganowsky, M. Landau, M. Zhao, A. B. Soriaga, L. Goldschmidt, D. Flot, D. Cascio, M. R. Sawaya, and D. Eisenberg, Proc. Natl. Acad. Sci. U.S.A. **108**, 16938 (2011).
[10]J. E. Straub and D. Thirumalai, Curr. Opin. Struct. Biol. **20**, 187 (2010).
[11]P. Neudecker, P. Robustelli, A. Cavalli, P. Walsh, P. Lundstroem, A. Zarrine-Afsar, S. Sharpe, M. Vendruscolo, and L. E. Kay, Science **336**, 362 (2012).
[12]B. Ma and R. Nussinov, Curr. Opin. Cell Biol. **10**, 445 (2006).
[13]C. Wu and J. E. Shea, Curr. Opin. Struct. Biol. **21**, 209 (2011).
[14]R. Pellarin, E. Guarnera, and A. Caflisch, J. Mol. Biol. **374**, 917 (2007).
[15]G. Bellesia and J.-E. Shea, J. Chem. Phys. **130**, 145103 (2009).
[16]H. Krobath, S. G. Estácio, P. F. N. Faísca, and E. I. Shakhnovich, J. Mol. Biol. **422**, 705 (2012).
[17]H. D. Nguyen and C. K. Hall, Biophys. J. **87**, 4122 (2004).
[18]A. V. Smith and C. K. Hall, J. Mol. Biol. **312**, 187 (2001).
[19]S. Auer, F. Meersman, C. M. Dobson, and M. Vendruscolo, PLOS Comput. Biol. **4**, e1000222 (2008).
[20]J. Zhang and M. Muthukumar, J. Chem. Phys. **130**, 035102 (2009).
[21]H. Taketomi, Y. Ueda, and N. Gō, Int. J. Pept. Protein Res. **7**, 445 (1975).
[22]R. D. Hills and C. L. Brooks, Int. J. Mol. Sci. **10**, 889 (2009).
[23]L. Prieto and A. Rey, J. Chem. Phys. **130**, 115101 (2009).
[24]J. Karanicolas and C. L. Brooks, J. Mol. Biol. **334**, 309 (2003).
[25]F. Ding, N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley, and E. I. Shakhnovich, J. Mol. Biol. **324**, 851 (2002).
[26]T. X. Hoang, A. Trovato, F. Seno, J. R. Banavar, and A. Maritan, Proc. Natl. Acad. Sci. U.S.A. **101**, 7960 (2004).
[27]S. Auer, C. M. Dobson, and M. Vendruscolo, HFSP J. **1**, 137 (2007).
[28]S. Brown, N. J. Fawzi, and T. Head-Gordon, Proc. Natl. Acad. Sci. U.S.A. **100**, 10712 (2003).
[29]J. W. Mullinax and W. G. Noid, Proc. Natl. Acad. Sci. U.S.A. **107**, 19867 (2010).
[30]M. Enciso and A. Rey, J. Chem. Phys. **136**, 215103 (2012).
[31]E. V. Koonin, Y. I. Wolf, and G. P. Karev, Nature (London) **420**, 218 (2002).
[32]M. Enciso and A. Rey, J. Chem. Phys. **132**, 235102 (2010).
[33]L. Prieto, D. de Sancho, and A. Rey, J. Chem. Phys. **123**, 154903 (2005).
[34]L. Prieto and A. Rey, J. Chem. Phys. **127**, 175101 (2007).
[35]M. Larriva, L. Prieto, P. Bruscolini, and A. Rey, Proteins **78**, 73 (2010).
[36]S. Auer, A. Trovato, and M. Vendruscolo, PLOS Comput. Biol. **5**, e1000458 (2009).
[37]D. Thirumalai, D. K. Klimov, and R. I. Dima, Curr. Opin. Struct. Biol. **13**, 146 (2003).
[38]M. J. Bennett, M. R. Sawaya, and D. Eisenberg, Structure **14**, 811 (2006).
[39]C. M. Dobson, Methods **34**, 4 (2004).
[40]C. Clementi, Curr. Opin. Struct. Biol. **18**, 10 (2008).
[41]Y. Liu and D. Eisenberg, Protein Sci. **11**, 1285 (2002).
[42]J. Li, C. L. Hoop, R. Kodali, V. N. Sivanandam, and P. C. van der Wel, J. Biol. Chem. **286**, 28988 (2011).
[43]J. King, C. Haase-Pettingell, A. S. Robinson, M. Speed, and A. Mitraki, FASEB J. **10**, 57 (1996); available online at http://www.fasebj.org/content/10/1/57.full.pdf+html.
[44]S. Yang, S. S. Cho, Y. Levy, M. S. Cheung, H. Levine, P. G. Wolynes, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **101**, 13786 (2004).
[45]A. Caflisch, Curr. Opin. Struct. Biol. **10**, 437 (2006).
[46]G. G. Tartaglia, A. P. Pawar, S. Campioni, C. M. Dobson, F. Chiti, and M. Vendruscolo, J. Mol. Biol. **380**, 425 (2008).
[47]E. Y. Chi, S. Cite, and T. W. Randolph, Pharm. Res. **20**, 1325 (2003).
[48]See supplementary material at http://dx.doi.org/10.1063/1.4820793 for the technical details on the clustering procedure used in this work, and for Figure S1, which shows the energy histograms from the single chain simulations along the folding transition temperatures.

Sketching protein aggregation with a physics-based toy model

Marta Enciso and Antonio Rey

SUPPLEMENTARY MATERIAL

Clustering procedure

The clustering analysis of the discussed simulations has been performed through an in-house program that reads and analyzes the trajectory files generated by the main program. In these trajectories, the coordinates of all the units of the simulated system are recorded in different files corresponding to the different temperatures sampled in the REMC procedure. From these coordinates, the analysis program counts the number and type of hydrogen bonds formed in each configuration, classifying them into helical or β-type. In the same way, hydrophobic "contacts" are also computed and identified as inter/intra chain. Note that the hydrogen bond interaction is a square-well potential in our model, while the latter follows a Lennard-Jones functional form. For this reason, we have considered that a hydrogen bond is formed if the geometric restrictions included in its definition are fulfilled (see Methods in the main manuscript for details) while a hydrophobic interaction is counted as a "contact" if it is attractive and its strength is at least 40% of its maximum value. Besides, the radius of gyration is computed for each chain and chain-to-chain distances (among their centers of mass) are also calculated.

Using all this information, we have first identified which of the abovementioned properties are characteristic of the native state (for each kind of sequence), mainly the number and type of hydrogen bonds and the radius of gyration. Thanks to this single-chain analysis, isolated folded chains are identified as "native" in the multichain simulations. Domain-swapped configurations present a number and type of hydrogen bonds and hydrophobic contacts comparable to the native conformations, but have larger radius of gyration (as well as lower chain-to-chain distances). "Aggregated" configurations present almost exclusively β-type hydrogen bonds and always involve two or more chains. Configurations that do not lie within these groups are classified as denatured. Following this procedure, every recorded configuration in the trajectory files is classified as belonging to one of the

classes mentioned above. Adding up the number of recorded configurations belonging to the different classes provides the populations we report in our results (for example, in Figure 2C) and which allow us to compute the structural phase diagrams in Figures 4, 5C and 6C.

After this numerical analysis, the different clusters are also visually inspected to further check the accuracy of the classification criteria and the variability within each of these clusters. This way, we have checked that the cartoons shown in Figures 2B, 5B and 6B are representative of the clusters to which they belong.

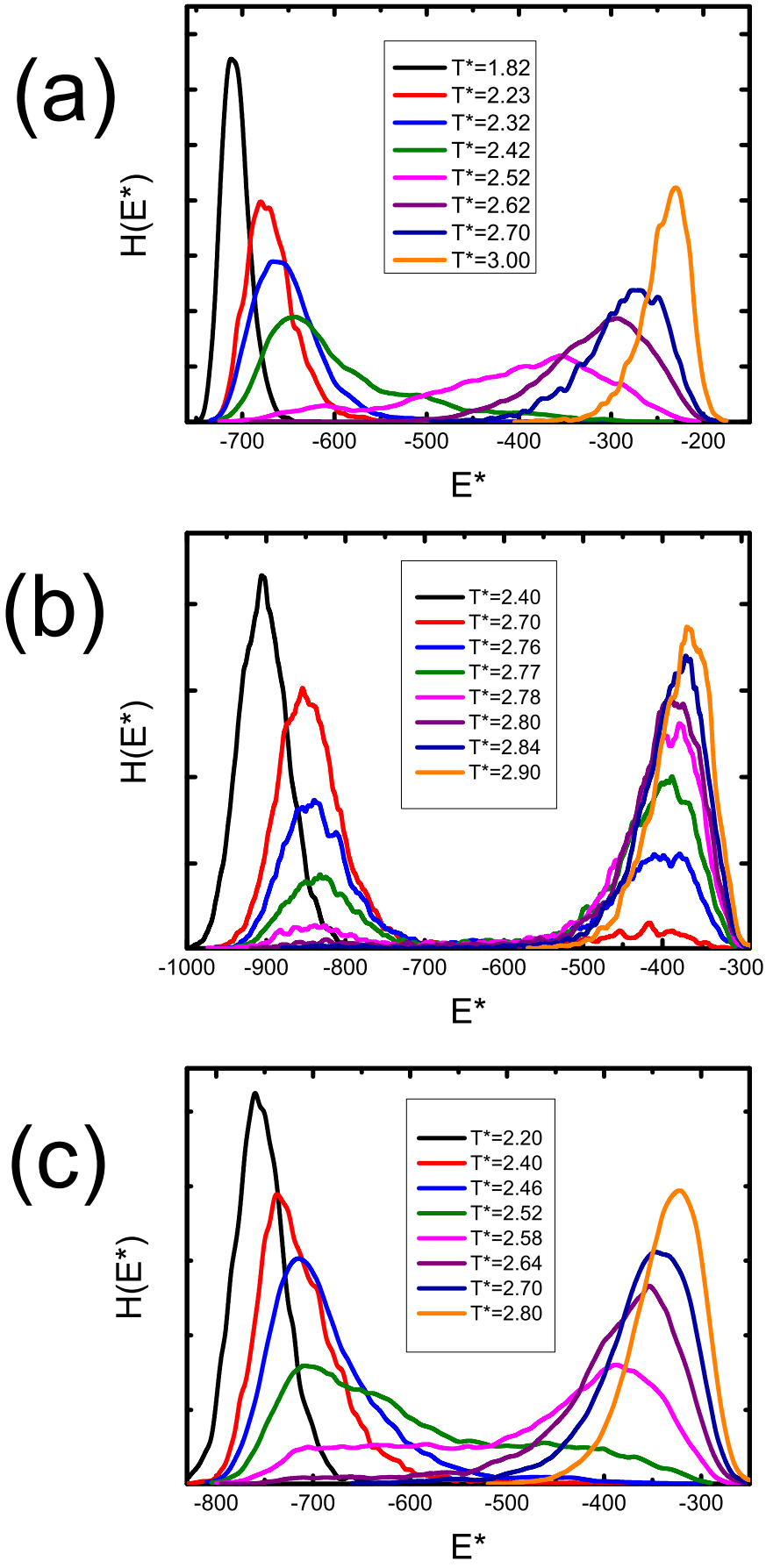


Figure S1. Energy histograms from the equilibrium simulations of isolated polypeptide chains at different selected temperatures around the transition region: (a) all-α protein; (b) all-β protein; (c) $\alpha + \beta$ protein.